Communication and Attitude Revision

Douglas E. Appelt

SRI International
333 Ravenswood Ave.
Menlo Park, California

1 Introduction

Much recent research has been directed toward understanding those aspects of language use that fall into that somewhat ill-defined area between semantics and pragmatics. The linguistic phenomena that seem to fall into this area include presupposition, implicature, speech acts (especially performatives), metonymy, and metaphor. These linguistic pheonmena can be characterized by a failure of truth conditional semantics alone to provide a satisfactory account, which is manifested in an obvious discrepancy between the "superficial" or "literal" content of the sentence and the intention underlying the speaker's use of the utterance in a particular situation.

Several theories have been evolving that are directed toward explaining these phenomena on the border between semantics and pragmatics, which could be characterized broadly as *update theories*. These include theories of mental state revision resulting from speech acts (Cohen and Levesque [5], Perrault [13], Appelt and Konolige [1]), update semantics (Heim [6], Zeevat and Scha (this volume), Thomason [16]), and abduction (Hobbs et al. [7], Hobbs (this volume), Charniak and Goldman [2]). All of these general frameworks have a common thread: a view of an utterance as an action that transforms an initial state of the world into a resulting state, and in the process producing a set of changes to the mental states of the participants. These changes are represented as an update to a model of their respective mental states. The meaning of the utterances in the most general sense is identified with the changes they produce in this model, rather than with their truth-conditional semantics alone.

Mental state revision models of speech acts adopt the perspective that utterances reveal constraints on the mental states of the participants. According to this view, the fact that a speaker utters a sentence constrains his mental state by default to conform to certain conditions that are consistent with the sentence's meaning and the intention to utter it, and these constraints become public knowledge as a result. Similarly, the hearer's mental state is affected in various ways as he adapts his beliefs and intentions to the information gained about those of the speaker. Implicit in all the variations of this approach is the

rejection of the a causal role for illocutionary acts in the belief revision process. Some of the earliest work on speech act planning (e.g., Cohen [4]) was predicated on the assumption that agents must explicitly recognize the illocutionary force of the utterance to respond to it appropriately. If someone says "It's cold in here," hoping that the hearer will shut the window, it was believed that the hearer would shut the window if and only if he recognized the fact that "It's cold in here" is a request.

According to mental state revision models, the revision of attitudes does not take place because an utterance has been recognized as a particular kind of speech act, but it is rather a rational response by the discourse participants to new information that is presented by the utterance. The sentence "It's cold in here," provides some information about the mental state of the speaker, and given principles of rationality and mutual belief, the hearer can conclude that the speaker must have other beliefs and intentions that are not realized directly in the utterance. The labeling of the act as a "request" is something a theorist might do in analyzing the situation, but the hearer need not do that to respond appropriately. Different illocutionary acts characterize different patterns of revision of mental state, but the revision is not a response to their recognition.

Update semantics shares much in common with the mental state revision models, except that the focus is on the updating of a shared conversational record rather than the complete constellation of private and public attitudes of the participants.

Abduction appears on the surface to be an entirely different approach to pragmatic interpretation, but it is in fact quite similar. Abduction can be viewed as a process of diagnosis, in which unobservable causes are inferred to account for observations. In the domain of sentence interpretation, observations consist of a literal semantic representation of the sentence uttered. This logical form is proved from a knowledge base representing the hearer's initial mental state, allowing the assumption of the key "unobservable" propositions needed to complete the proof. The minimal set of assumptions that permit the observation to be proved can be considered the incremental update to the hearer's beleifs.

In this paper I examine two frameworks for mental-state update theories, a variant of autoepistemic logic, and weighted abduction, to determine their relationships and determine the suitability of each for stating a general theory of semantics and pragmatics.

2 In Search of Literal Meaning

The notion that sentences have a literal meaning independent of their context (or one in which the context can be incorporated straightforwardly as a "variable") is somewhat controversial. This issue is particularly important for update theories, since it is a central feature of these theories that updating of a mental state model proceeds from some representation of a literal content of the utterance.

The main problem that a notion of literal meaning must confront is the problem of real semantic ambiguity. Certain sentences, even those as simple as

"Every man loves a woman," have a space of possible meanings that depend on the assignment of scope to various operators, completely independent of whether any predicates in the sentences are used in a literal, metonymic, or metaphorical sense. For all the scoping possibilities in a sencence like "Every man loves a woman," which of them is to be considered the representation of its literal, context independent content? A similar problem results from lexical ambiguity. In a sentence like "The secret agent hid the microfilm in the pen," the predicate "pen" could refer to a writing instrument, or a fenced enclosure. Which of these possibilities is the "literal" interpretation?

The notion of literal meaning is not inconsistent with ambiguity, although solving the problem of representing the ambiguities that do arise can be difficult. In the worse case, the literal meaning is a disjunction of possible ambiguous interpretations, but the explosion of possibilities rules out this representation for all but a few types of ambiguity. A better solution is to find representations that are vague among several more specific interpretations. This solution is supported by the fact that people are in general not aware of ambiguous alternatives without making an explicit effort to uncover ambiguities (e.g., Van Lehn [17]). Hobbs' approach to the representation of quantifiers (Hobbs, [8]) is an example of how this principle of vagueness can be applied to the problem of quantifier scoping. A semantically neutral interpretation of a lexically ambiguous word can be obtained by using a predicate that generalizes over the space of possible lexical ambiguities.

It is important to realize that the truth-conditional interpretation of the literal content does not have to be consistent with *any* context-dependent update of the discourse participants mental states. For example, just because the sentence "This meal was delicious" could be interpreted ironically in certain contexts does not mean that its literal meaning has to be vague with respect to a proposition and its negation. If that were the case, the literal interpretation of most sentences would be so vague as to be consistent with almost anything.

The attractiveness of update theories for accounting for certain phenomena like irony and metaphor is that the updated attitude model does not necessarily have to be truth-conditionally consistent with the literal content of the utterance, nor does this update have to take place in a single step. The precise specification of literal meanings and the inferences that are drawn from them to update a attitude model are the meat of the fields of semantics and pragmatics, and it is of course impossible to discuss all of the problems. However, the rejection of the possiblity of determining a literal logical form for an utterance is at this time, fortunately, quite premature.

3 A Logic for Expressing Facts about Attitude Revision

Appelt and Konolige have proposed a theory of speech acts based on Hierarchic Autoepistemic Logic (HAEL) [1]. This work attempts to provide a theory of the changes produced in the mental states of participants in a dialogue from the standpoint of an observer. The observer's theory contains partial information

about each agent's theory of the world, and the utterance produces updates to both the speaker's and hearer's theories.

Autoepistemic (AE) Logic was originally formulated by Moore [12] to address some difficulties in McDermott's formulation of nonmonotonic modal logics [11]. The central idea is to devise a formalism that can capture an agent's ability to draw conclusions that follow from completeness or incompleteness of his own knowledge. A typical example would be answering the question of whether I have an older brother. Although I can't *prove* that I don't have one, I can reason that if I did, I would certainly know about it. Since I do not know that I have one, I therefore conclude that I do not.

Standard AE logic consists of a first-order theory augmented with a modal operator, L. Sentences of AE theory T, in addition to any first order sentences, can contain sentences of the form $L\phi$, which is interpreted as "ϕ is a theorem of T." A stable expansion E of an AE theory is a set of sentences that satisfies the following conditions:

1. $T \subseteq E$
2. E is closed under first order consequence
3. if $\phi \in E$, then $L\phi \in E$
4. if $\phi \notin E$, then $\neg L\phi \in E$

The intuitive interpretation of the L operator is self knowledge. If $L\phi$ is a theorem, one can say that "The cognizing agent believes that he believes ϕ." Similarly, $\neg L\phi$ is intuitively interpreted as "The cognizing agent believes that he doesn't know whether ϕ is true."

A stable expansion is a maximal set of consistent sentences that can be derived from the original theory. Because axioms can be applied in different sequences to derive consequences that may be mutually inconsistent, an autoepistemic theory will, in general, have multiple stable expansions, each representing some alternative way of drawing a consistent set of consequences from the base theory. These stable expansions are like the extensions of a default logic [15]. Konolige [9] demonstrated the formal equivalence between autoepistemic and default logic. Therefore, Perrault's [13] default logic formulation of speech act theory can be mapped straightforwardly into an autoepistemic formalism with equivalent representational power.

The fact that an AE theory can have multiple stable expansions presents some problems if we wish to use the theory to predict consequences of events. Does the theory predict something if it holds in *some* stable expansion? If so, and the theory has multiple stable expansions, then it makes inconsistent predictions. Does it predict something if it holds in *every* stable expansion? In that case, the predictive power of the theory may be too weak to account for the facts we want to explain.

What we would like is a theory that offers the theorist the possibility of excercising more control over conflicting defaults. It is with this idea in mind that Hierarchic Autoepistemic Logic was devised.

AE logic is extended to HAEL by decomposing the theory T into a number of subtheories $T_i, T_j, \ldots$ together with a partial order $\prec$ on these theories. If

$T_i \prec T_j$, then every theorem of T_j is also a theorem of T_i. Instead of a single L operator, an operator L_i is introduced for each T_i, and the entire theory is subject to the constraint that, if L_j occurs positively (negatively) in T_i, then $T_j \preceq T_i$ ($T_j \prec T_i$). The definition of a stable expansion is suitably modified to take into account the multiple theories, and the constraints of which operators can refer to which theories.

This decomposition of an AE theory into a hierarchy of subtheories gives us the capability of representing the strength of an agent's beliefs. We assume that an agent's beliefs consist of all the facts that hold in some maximal level of the hierarchy. Lower levels of the hierarchy represent strongly held beliefs, while higher levels represent progressively weaker beliefs. The rules at each level can be used to conclude weaker beliefs based on the presence or absence of stronger beliefs at the lower levels. In describing the persistence of beliefs across state transitions, it is easy to state that beliefs at the higher, weaker levels persist subject to lack of contradiction by beliefs that hold at lower, stronger levels.

To formalize the effects of speech acts within this theory, we assume that the HAEL theory is augmented with modal operators for each pair of agents and states for representing beliefs and intentions of agents in that state. We use the notation "$[a_i]\phi$" to represent agent a's belief in state i that ϕ is true. Similarly, we represent intention with an indexed set of modal operators. The formula $\{a_i\}\phi$ means that a intends in state i to bring it about that ϕ is true. We furthermore assume that propositions ϕ are true with respect to particular states, and that the logic includes a temporal operator $\Box\phi$ meaning that ϕ is true in all possible future states.

The specific semantics of the modal operators is not of central importance to the theory. Any reasonable definition for $[a_i]$, such as an $S4$ logic, is acceptable. Similarly, any definition of $\{a_i\}$ is acceptable as long as it obeys the following properties. The first property is ***belief introspection***, i.e. agents are aware of their own intentions.

$$\vdash \{a_i\}\phi \supset [a]\{a_i\}\phi.$$

Furthermore, agents have consistent intentions, i.e.,

$$\vdash \{a_i\}\phi \supset \neg\{a_i\}\neg\phi,$$

and they do not intend what they believe to be impossible:

$$\vdash [a_i]\Box\neg\phi \supset \neg\{a_i\}\phi,$$

or inevitable:

$$\vdash [a_i]\Box\phi \supset \neg\{a_i\}\phi.$$

The fundamental idea that makes HAEL useful for applications in reasoning about attitude revision is that an agent's beliefs are represented not by a monolithic theory, but by the union of the beliefs represented in a number of theories, ordered in a hierarchy so that the relative strength of believes in the propositions is proportional to the level of the hierarchy in which they hold. Strong beliefs are represented by the theories in the lower level of the hierarchy, weaker beliefs are represented by higher levels.

This hierarchy gives one a handle on the problem of representing the persistence of beliefs across state. Typically, a belief persistence axiom describing the relationship between an agent a's beliefs in an initial and final state is a nonmonotonic axiom schema of the form
In theory T_i:

$$[a_0]\phi \wedge \neg L_{i-1}\neg[a_1]\phi \supset [a_1]\phi.$$

This rule says that agent a's belief that ϕ persists from state 0 to state 1 in theory T_i as long as nothing provable at level $i-1$ contradicts it.

An important point is that this update rule can be stated without any reference to specific rules in theory T_{i-1}. It's advantage is that the prioritization of *theories* rather than *rules* makes it possible to have a theory of belief persistence that is independent of the particular rule formulation that is chosen to express those beliefs.

An advantage of default or autoepistemic logics for the formalization of the effects of speech acts is that it is possible to make a very concise characterization of the effects of speech acts that agrees well with one's intuitions, and seems to make the correct predictions about the effects of speech acts, given some relatively straightforward assumptions about belief revision.

The HAEL-based speech-act theory is assumed to include an "utterance" theory, u, that reflects the literal meaning P of the sentence uttered. A speaker s utters a sentence with semantic content P in an initial situation i, resulting in a final situation f. The utterance theory contains the semantic content of the utterance, pus a set of carefully delimited rules from which a set of propositions is derived that constitute what is "up for consideration" as a result of the speaker uttering the sentence. This provides the base upon which inferences are performed to deduce the mental states of the dialogue participants in the situation f, given certain information about what they believe in situation i.

The HAEL speech act theory consists of rules that relate the contents of the utterance theory to the beliefs of the speaker and the hearer in states i and f. The determination of the speaker's and hearer's mental states in situation f can be thought of as analogous to a database update, with the contents of u providing the basis for that update.

In addition to containing the semantic representation of the declarative utterance, P, we assume that u contains the schemata

$$[u]\phi \supset [u]\{s_i\}[h_f]\phi$$

and

$$[u]\phi \supset [u]\{s_i\}[h_f][s_f]\phi.$$

This says that an utterance brings into consideration not just its meaning, but also the speaker's intention that the hearer believe that meaning (a perlocutionary intention), and the speaker's intention that the hearer believe the speaker believes it (an illocutionary intention).

The notion of "update" is captured by relating the speakers and hearer's beliefs to the contents of the utterance theory in the state f resulting from the utterance.
In T_1:

$$[u]\phi \wedge \neg L_0 \neg [s_f]\phi \wedge \neg L_0 \neg \{s_i\}[h_f]\phi \supset [s_f]\phi$$

$$[u]\phi \wedge \neg L_0 \neg [h_f]\phi \wedge \neg L_0 [h_f] \neg [s_f]\phi \wedge \neg L_0 [h_f] \neg \{s_i\}[h_f]\phi \supset [h_f]\phi$$

The first axiom describes how an utterance constrains the speaker's beliefs. We conclude that the speaker believes what he says as long as it is consistent with his strongly held beliefs ($\neg L_0 \neg [s_f]\phi$) and he actually intends that the hearer believe it ($L_0 \neg \{s_i\}[h_f]\phi$). The hearer believes the proposition expressed as long as it is consistent with his strongly held beliefs ($\neg L_0 \neg [h_f]\phi$), it is consistent that the speaker believes what he says ($\neg L_0 \neg [h_f] \neg [s_f]\phi$) and it is consistent that the speaker is using the utterance with communicative intent ($\neg L_0 [h_f] \neg \{s_i\}[h_f]\phi$).

This axiom is the first instance of a schema extending the conclusion to mutual belief. Other instances of this schema match the above axioms, with progressively deeper nesting of one agent's belief about the other at each level.

This account of speech acts shares some similarities with Perrault's [13] theory based on normal default logic, and certainly many of its theoretical motivations. One difference is the use of an "utterance theory" to capture a collection of propositions that are brought into consideration by the utterance of a particular speech act. Under this view, not only the literal proposition of the speech act, but certain systematically related propositions are also considered relevant to the belief and intention revision process.

The most important difference with Perrault's formulation centers on the treatment of belief revision. Perrault assumed (for the sake of argument) that beliefs always persist from one state to the next, and therefore defeat attempted utterances that contradict them. Also, agents would remember their beliefs, from one state to the next, however, they would immediately forget anything about their ignorance.

Although Perrault's belief revision "theory" was certainly intended as a simplification of reality, it is not clear how to remedy the defects within the default logic framework he initially proposed. It is clear that the formalism must account for the persistence of ignorance from one state to another. Otherwise, the theory would predict that agents could convince themselves of something they didn't believe merely by asserting it. But, a default rule that concludes the persistence of ignorance from one state to another will create multiple extensions with respect to belief in the proposition uttered. As a theory of speech acts, this consequence of multiple extensions is undesireable, because unless one can formulate clear criteria for which extension is prefered, the theory taken as a whole makes no interesting predictions about the agents' beliefs. All the theory can tell you is that the speech act has certain effects, or it does not.

It is well known that one can prioritize defaults in a default theory by transforming normal defaults to non-normal defaults. The strategy is to add conditions to the antecedents of rules that block the application of the default rule in

situations in which the conclusion derived by another rule is derived first. For example, if

$$p \wedge \neg L \neg q \supset q$$

is one rule in an autoepistemic theory and if

$$r \wedge \neg L q \supset \neg q$$

is another rule, the theory will have stable expansions corresponding to the consequences of each of the rules. If we wanted to prioritize these two rules, we could add conditions to the second rule to prevent it from applying in any situation in which the first rule applies. One way to do this would be to reformulate the second rule as

$$r \wedge \neg L(q \vee p) \supset \neg q.$$

It is easy to see that to express priorities correctly, the default rules must be analyzed to determine their dependencies, and these dependencies expressed as additional conditions on the default rule applicability. This solution to the belief revision problem is undesireable because in addition to being difficult, if not impossible to do, it precludes the possibility of separating knowledge about speech acts from a model about agents' beliefs in general. Under this approach one could not claim that a theory of communication is one aspect of commensense knowledge that all agents share, because every agent's theory would depend on his belief revision strategy, and hence be different.

The HAEL theory makes it possible to state theoretical predictions about each agent's belief revision processes by allowing one to hypothesize relative strengths of beliefs in propositions. Because every consistent HAEL theory has exactly one stable expansion (Konolige, [10]) propositions about an agent's mental state become a well-founded consequence of the theory, rather than the consequence of the theory with respect to some extension. Furthermore, under this model, the description of the effects of speech acts is the same for all agents. The fact that the same speech act can have different effects on different agents is a consequence of the differing beliefs and belief revision strategies.

4 Default Theories and Abductive Interpretation

A theory of mental state revision cast in terms of prioritized defaults like the HAEL theory discussed in the previous section, as well as an update semantics theory (Zeevat and Scha, this volume) can be thought of as a constructive process, in which the content of an utterance is used together with a description of the participants' current state to compute the characteristics of a new state that accommodates the new information to the previous state. In a theory of abductive interpretation (Hobbs et al., [7]), the process is viewed somewhat differently. The content of the new utterance is taken as a fact to be explained. This explanation is accomplished by adding assumptions to a theory of the initial state which would allow the derivation of the content of the utterance from the initial theory plus the assumptions. If these assumptions become a permanent part of

the theory for the interpretation of subsequent utterances, the assumptions can be considered an "update" to the initial theory. Preferences among alternative updates explaining the same utterance are indicated by weighting factors on the antecedent literals of the theory's rules.

Update theories of all varieties are confronted with a problem when the intended interpretation of the utterance is logically inconsistent with its literal interpretation, or involves the flouting of Gricean conversational maxims, such as in the cases of irony and metaphor. Although the details of a theory of metaphor within the framework of HAEL remain to be worked out at this time, a likely approach within this framework would be to define several update strategies from the same utterance, some depending on a literal interpretation, and others depending on a systematically derived metaphorical interpretation. However, any conclusions derivable from the metaphorical interpretations would be defeated by any conclusion resulting from the literal illocutionary or perlocationary intentions of the utterance. For example, if a declarative utterance is strongly mutually believed to be false, then all conclusions about illocutionary and perlocutionary intentions of the speaker relating to its belief by the hearer must be defeated (assuming the theory that obeys the belief and intention constraints outlined in the previous section) and the literal utterance has no effects. Only in such a case would the effects derivable from one of the metaphorical interpretations be adopted.

Hobbs (this volume) proposes a quite different analysis of metaphor within the framework of abductive interpretation that treats metaphor similar to metonymy. The goal is to explain the meaning of the speaker's utterance (which ordinarily would be unexplainable in the case that the literal interpretation of a metaphorical utterance would be mutually believed to be false) by finding a systematically related interpretation that *can* be explained. Interpretation of metonymy relies on finding an individual systematically related to the individual actually satisfying the description in the utterance, but which satisfies constraints on the types of individuals that can participate in the relations posited by the sentence. Interpretation of metaphor involves weakening or transforming the posited relations so that they hold for the same set of individuals referred to in the utterance. To accomplish this, Hobbs proposes a meta theory, which maps the axioms of the primary theory into axioms of a metaphor interpretation theory that can consistently hold of the mentioned set of individuals.

The proposal has at least one serious problem. To provide an adequate account of metaphor, the account of interpretation as abduction [7] needs to be revised or generalized. The characterization of pragmatic interpretation as an abductive proof of the sentence's truth is very elegant but its elegance relies on certain assumptions about the communicative situation. Abductive interpretation is most appropriate for texts like newspaper articles (although newspaper articles contain metaphors and the problem does not go away) for which the reader has no specific knowledge of who the speaker is, or details about his particular mental state. This justifies assuming that the speaker is sincere and identifying mutual knowledge with the projection of the interpreter's own knowledge, and it justifies identifying abductive assumptions with the augmentation

of the reader's or hearer's own knowledge.

These simplifying assumptions are not justified if one wants to account for the interpretation of metaphor, because in many cases metaphor involves transparent insincerity. The most serious problem, however, is that some utterances whose interpretation is intended metaphorically are literally true. An example of such a metaphor is the utterance of "I am not Donald Trump" in response to a request to borrow a large sum of money. This utterance works like a metaphor, because the literal content of the utterance is to be taken metaphorically as a denial of one of Donald Trump's contextually salient properties (which until recently was his great wealth). This interpretation of the metaphor then provides the input for the conversational implicature that the utterance constitutes a denial of the request. The problem with the standard strategy of abductive interpretation as outlined by Hobbs [7] is that the mutual belief that an utterance is true *is always* the best explanation of its truth taken in isolation. No assumptions are necessary, and therefore the minimal accommodating update is no update at all! Similarly, some metaphorical statements do not violate any selectional constraints, as in "He hit the ball out of the park" in response to a question about how John did with his Ph.D. orals.

Rejecting a literal interpretation of the sentence in favor of its metaphorical one can sometimes be accomplished if the proof of the metaphorical interpretation accounts for discourse coherence as well as its own truth. In the Donald Trump example, the correct metaphorical interpretation is the only one that can generate an implicature that either assents to or denies the request. The problem with exclusive reliance on this mechanism is that the tendency to favor the assumption-free literal interpretation creates a very strong presumption that must be overcome. It is not at all clear that a consistent set of preferences can be devised based on explaining discourse coherence that will be sufficient to overcome this presumption in all cases.

If we seek to maintain the view that interpretation is the explanation of *something* and that the resulting augmentation of the interpreter's knowledge is considered to be those assumptions required for the explanation, then we need to change the specification of what it is that is being explained by an interpretation. An obvious candidate is the observation that the speaker uttered the sentence. The utterance of the sentence would be explained by hypothesizing the speaker's intended interpretation, and a set of mental attitudes that would support his intention to utter a sentence with that interpretation. Thus, the literal interpretation of a literally true metaphor could never be intended by the speaker (assuming its truth is mutually believed), because the assumption of such an intention would be inconsistent with the principle that agents do not adopt intentions to achieve what they already believe to be the case.

The explanation of the observation "Speaker utters S" is done by an argument supported by assumptions or facts of the form "S means ϕ" and the speaker's beliefs and intentions with respect to ϕ. Grice's maxims are reflected in the preference rules that one uses to judge one explanation better than another. For example, the maxim of quality could be expressed as a preference for explanations that assume the speaker's beliefs are consistent with the hearer's

beliefs, as opposed to assuming beliefs that are inconsistent. The maxim of relevance can be analogously stated in terms of preferences among explanations. For example, if the speaker's goal is to answer the hearer's previous question, then any explanation that does not contribute to that goal is assigned a very low evaluation.

One could start writing formal rules within the weighted abduction formalism that would allow conclusion of "Speaker utters S" from assumable premises. For example, one such rule might be something like Rule 1:

$$\text{Meaning}(u, \phi)^{(\alpha_1)} \wedge \{s\}[h]\phi^{(\beta_1)} \wedge [s]\phi^{(\gamma_1)} \supset \text{Utter}(s, u)$$

This rule expresses the intuition that if utterance u means ϕ, and the speaker intends that the hearer believe ϕ and the speaker believes ϕ, then the speaker utters u. The assumption weights reflect the fact that if the antecedents are consistent with the hearer's beliefs but not provable, they can be assumed at some cost proportional to the weight.

In a simple case, upon hearing an utterance u, a hearer would typically take its literal interpretation ϕ as the proposition expressed, and would assume that the speaker believes it and that he intends the hearer to believe it as well. Of course, this abductive formulation does not augment the theory with respect to new beliefs of the hearer. All the explanatory assumptions deal with the speaker's mental state. Thus in this more general case, abduction only tells part of the speech act story. The hearer must still decide what to believe.

Rule 1 makes a major omission, however. The conditions represented on the left hand side of the rule are neither necessary nor sufficient conditions for the speaker to utter u. For example, the speaker may not believe ϕ at all; he may be lying. The above rule cannot explain a lie, because it's premises are always inconsistent with a mental state model in which the speaker believes $\neg\phi$. To handle lies, we need another rule, something like Rule 2:

$$\text{Meaning}(u, \phi)^{(\alpha_2)} \wedge \{s\}[h]\phi^{(\beta_2)} \wedge [s]\neg\phi^{(\gamma_2)} \supset \text{Utter}(s, u).$$

Presumably the assumption weighting factors are different for this rule, in which γ_1 is much less than γ_2, reflecting the intuition that explanations assuming sincerity are better than explanations assuming insincerity given no other information to bias the conclusion.

This rule however, does not completely account for all the possibilites. To cite some of the omissions, it doesn't explain certain insincere utterances in which the literal content is consistent with the speaker's and hearer's beliefs. but it can be concluded from other evidence that the speaker's communicative intention is absent, and it doesn't distinguish confirmations from assertions. The problem is that the truth or falsehood of any one of an infinite number of propositions about the speaker's mental state potentially provides evidence for an explanation of why he would utter u.

5 Problems with Abductive Inference of Attitudes

The rules for "explaining" utterances as consequences of beliefs and intentions of the speaker, as outlined in the previous section, have at least two problems. The first problem results from the fact that the rules given are not *causal* rules. Two rules concluding the performance of a speech act from the intention to utter it and the belief that it is true (Rule 1) on one hand and the belief that it is false on the other (Rule 2) are logically redundant: it is possible to express these two rules as one rule that depends only on the intention to utter the sentence. The introduction to the theory of Rule 1 and Rule 2 was necessitated only by the need to state a preference for $[s]\phi$ over $[s]\neg\phi$.

The other problem, which is much more serious, is that abductive theories implicitly assume that the set of relevant explanatory hypotheses is closed. It is this assumption of closure that licences the inference of P from the rule $P \supset Q$ and the observation Q. If P is the only condition we know that is associated with Q, then the observation of Q at least strongly suggests P. But in the case of mental attitudes and speech acts, we *know* that the set of explanatory hypotheses about the speaker's mental state is not finite, at least if we take arguments about mutual belief seriously (e.g., Clark and Marshall [3], Perrault and Cohen [14]).

This assumption about closure is a reason to prefer a theory of mental state update based on a default or autoepistemic logic approach such as the HAEL theory outlined above, which starts from a closed set of hypotheses about the utterance's meaning, and implicitly characterizes an infinite number of potential conclusions about the speaker's and hearer's mental state, to an abductive theory, which requires the inclusion of all possible hypotheses about an agent's beliefs and intentions as potential assumptions from which the fact of uttering a sentence with the right propositional content is derived as a conclusion.

This argument, of course, does not deny the usefulness of the abductive approach as a computational tool to be applied in circumstances where a closure hypothesis on mental state assumptions can be reasonably applied. However, a competence theory of speech acts should account for the multiplicity of uses to which speech acts can be put, not just the "usual" or "normal" cases. That means that the theory should in principle account for all of the infinite effects that an utterance can have on the participant's mental states, and this appears to be impossible to do within an abductive theory.

6 Conclusion

In this article I have discussed two primary observations: (1) A proper account and characterization of speech acts, including an analysis of performatives, requires a theory to address the effect of the utterance on private as well as public attitudes, (2) Speech acts can produce an infinite number of potentially defeasible effects on the participants' mental states. A challenge to the successful formulation of a theory of speech acts is to finitely represent these multiple effects in a way that captures the agents' competence in a computationally usable

theory. Default-based theories have the advantage of being able to concisely state the effects of utterances in an intuitively plausible manner. Formulating the same facts within an abductive theory is not possible because the space of possible assumptions is not closed.

The particular default theory based on HAEL accomplishes this in a logical theory in which a reasonable model of belief revision and persistence can be stated. I believe its satisfaction of these criteria makes a strong case for further research using the nonmonotonic logic paradigm and HAEL in particular as a theoretical framework for developing more detailed theories of speech acts that correctly account for problems such as metaphor, irony, presupposition and implicature.

References

1. Appelt, D., Konologie, K.: A practical nonmonotonic theory for reasoning about speech acts. In *Proceedings of the 26th Annual Meeting.* Association for Computational Linguistics, (1988) 170–178
2. Charniak, E., Goldman, R.: A semantics for probabilistic quantifier-free first-order languages with particular application to story understanding. In *Proceedings of the 11th International Joint Conference on Aritificial Intelligence*, (1989) 1074–1079
3. Clark, H., Marshal, C.: Definite reference and mutual knowledge. In Joshi, A., Sag, I.. and Webber, B., *Elements of Discourse Understanding.* Cambridge University Press, Cambridge, England (1978)
4. Cohen, P.: *On Knowing What to Say: Planning Speech Acts* Ph.D. Thesis, Department of Computer Science, University of Toronto (1978)
5. Cohen, P., Levesque, H.: Rational interaction as the basis for communication. In Cohen, P., Morgan, J., and Pollack, M., *Intentions in Communication.* MIT Press, Cambridge, Massachusetts (1990)
6. Heim, I.: File change semantics and the familiarity theory of definiteness. In Bäuerle et al., *Formal Methods in the Study of Language.* Walther de Gruyter, Berlin, Germany (1983)
7. Hobbs, J., Stickel, M., Martin, P., Edwards, D.: Interpretation as abduction. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics* (1988) 95–103
8. Hobbs, J.: An improper treatment of quantification in ordinary english. In *Proceedings of the 21st Annual Meeting of the ACL* (1983) 57–63
9. Konolige, K.: On the relation between default and autoepistemic logic. *Artificial Intelligence* 35(3) (1988) 343–382
10. Konolige, K.: Hierarchic autoepistemic theories for nonmonotonic reasoning: Preliminary report. In Reinfrank, M., DeKleer, J., Ginsberg, M., Sandewall, E., *Non-Monotonic Reasoning*, Springer Verlag, Berlin, Germany (1989) 42–59
11. McDermott, D.: Nonmonotonic logic II: Nonmonotonic modal theories. *Journal of the Association for Computing Machinery*, 29(1) (1982) 33–57
12. Moore, R.: Semantical considerations on nonmonotonic logic. *Artificial Intelligence* 25(1) (1985) 75–94
13. Perrault, R.: An application of default logic to speech act theory. In Cohen, P., Morgan, J., and Pollack, M., *Intention and Communication.* MIT Press, Cambridge, Massachusetts (1990)

14. Perrault, R., Cohen, P.: Inaccurate reference. In Joshi, A., editor, *Formalizing Discourse*. Cambridge University Press, Cambridge, England (1980)
15. Raymond Reiter. A logic for default reasoning. *Artificial Intelligence* 13 (1980) 81–132
16. Richmond Thomason. Accommodation, meaning, and implicature. In Cohen, P., Morgan, J., and Pollack, M., *Intention and Communication*, pages 325–363. MIT Press, Cambridge, (1990)
17. Kurt VanLehn. Determining the scope of english quantifiers. technical report AI-TR-483, Massichusetts Institute of Technology Artificial Intelligence Laboratory (1978)

This article was processed using the LaTeX macro package with ICM style